University of St Andrews

CS5014 Coursework 1

Machine Learning

Author:
150008022

February 17, 2019

Goal

The goal of this practical was to cleanse and process real world data in order to produce a regression model, and evaluate its performance. [1]

Part I

Loading and Cleaning the Data

Numpy was used to load the csv file. The header row was skipped, and to ensure there were no missing values, the invalid raise flag was also used when parsing the data. This would raise an exception if any rows were found to be missing data.

The input and output columns were separated into two variables x and y, as in accordance with the notation used in lectures.

Part II

Analysing and Visualising the Data

Firstly, a histogram of each of the input variables and outputs were plotted in order to visualise the distribution of the values. When comparing to the histograms from the given paper TODO fig 1) ref paper, most of the plots matched. The difference was identified to be caused by 10 bins always being used (the default if not specified by numpy.hist), and the paper would sometimes use more. However, it was still clear that none of the variables had a gaussian distribution.

In order to identify which variables affected the outputs most, scatter graphs were made between each input and output variable. Feature scaling was used to allow for comparisons between values that could have very different ranges. Mean normalisation was used for the scatter plots. The resulting plots resembled those from the given paper (Figure 2 from given paper TODO), with different x-axis values since mean normalisation gave values in the range $-0.5 \leq x \leq 0.5$.

Conclusion

References

[1] Facebook. create-react-app.